Comparison of four species-delimitation methods applied to a DNA barcode data set of insect larvae for use in routine bioassessment

Bryan P. White[1], Erik M. Pilgrim[2], Laura M. Boykin[3], Eric D. Stein[1] and Raphael D. Mazor[1]

Abstract

Species delimitation (the process of grouping individuals into distinct taxonomic groups) is an essential part of evolutionary, conservation, and molecular ecology. Deoxyribonucleic acid (DNA) barcodes, short fragments of the cytochrome *c* oxidase subunit I (COI) gene, are being used in environmental bioassessments to assign specimens to putative species. However, no method for delimiting DNA barcodes into species-level entities has been universally accepted. We investigated the effect of delimitation methods on outcomes of bioassessments based on DNA barcodes. We applied two tree-construction methods (neighbor joining [NJ] and maximum likelihood [ML]) and 4 classes of species-delimitation criteria (distance-based, bootstrap support, reciprocal monophyly, and coalescent-based) to a DNA barcode data set consisting of three genera and 2202 COI sequences. We compared sets of species delimitations produced with different methods for *Baetis* (Ephemeroptera:Baetidae), *Eukiefferiella* (Diptera:Chironomidae), and *Simulium* (Diptera:Simuliidae) in samples from different streams. We assessed congruence among trees and compared species abundances and estimated species richness among methods. NJ followed by application of a standard barcoding distance cutoff (2%) resulted in the greatest number of putative species. All other delimitation methods yielded similar, but lower, species richness. Differences in species delimitations produced by various methods might have been caused by confounding factors, such as possible parthenogenesis in *Baetis* and rare haplotypes in abundant species of *Baetis* and *Simulium*. *Eukiefferiella* presented the fewest discrepancies among delimitations. Each method can be regarded as producing a separate line of evidence contributing to the delimitation of separately evolving lineages. The increased resolution offered by DNA barcoding can yield important insights into the natural history of organisms, but the power of these observations are limited without the use of multigene and multilocus data sets.

Introduction

DNA barcoding is a molecular taxonomic method wherein a ~650 basepair (bp) region of the cytochrome *c* oxidase subunit I (COI) mitochondrial

[1]*Southern California Coastal Water Research Project, Costa Mesa, CA*

[2]*US Environmental Protection Agency, Ecological Exposure Research Division, National Exposure Research Laboratory, Cincinnati, OH*

[3]*University of Western Australia, ARC Centre of Excellence, Plant Energy Biology, Western Australia, Australia*

gene is used as a species-level molecular identification marker in animals (Hebert *et al.* 2003). DNA barcoding has the potential to affect the field of bioassessment (also called biomonitoring) where biological measures, such as species richness and taxonomic composition, are used to draw conclusions about the health of an ecological system. Incorporation of DNA barcode methods into bioassessment programs has been evaluated theoretically (Jones 2008) and empirically (Sweeney *et al.* 2011; Pilgrim *et al.* 2011; Stein *et al.* 2013, In press). The conclusion is that molecular methods might improve our view of stream diversity because they provide increased taxonomic resolution.

In most bioassessment programs, estimates of species richness are obtained via morphological identification. However, this method can be difficult and time-consuming, and morphology cannot be used to differentiate cryptic species (Bickford *et al.* 2007) or species complexes (Hajibabaei *et al.* 2006). The ability to differentiate such species is important because presently unknown differences in their tolerance to pollution, reproductive timing, feeding mechanisms, or other ecological traits may provide clues regarding the health of a stream system (Verberk *et al.* 2013). Moreover, species-delimitation methods are relevant to more than the outcomes of routine bioassessment programs. They may have utility in the fields of criminal wildlife forensics (Dawnay 2007), biodiversity indexing (Janzen *et al.* 2009), detection of fish-market replacements (Maralit *et al.* 2013), ecology (Valentini *et al.* 2009), and biosecurity (Boykin *et al.* 2012). In each application, routine, repeated sampling of organisms should yield consistent sets of species designations across different laboratories.

Use of molecular taxonomic methods, such as DNA barcoding, to identify unknown organisms or organisms impractical to identify to species level on a routine basis, has led to questions related to how to delimit species based only on a DNA barcode. Rapid decreases in the cost of high-throughput Sanger sequencing and the advent of Next Generation Sequencing (NGS; Shendure and Ji 2008) have led to an exponential increase in the rate at which DNA barcodes are being uploaded to public databases. Thus far, 1.2 million COI sequences have been released on the Barcode of Life Data System (BOLD; Ratnasingham and Hebert 2007; www.barcodinglife.org,) as part of the International Barcode of Life Project (iBOL, http://ibol.org/), and another 1 million remain uploaded but unreleased. This vast amount of data has created a pressing need for efficient, objective, and readily reproducible algorithms for species delimitation. Many proponents of DNA barcoding have suggested the use of measures of genetic distance to designate species. Investigators have recommended species limits based on an average genetic distance of ≥2% (see below) among individuals in different putative species (Ball *et al.* 2005, Zhou *et al.* 2009) or on a level of interspecific variation that is 10× the intraspecific variation (Hebert *et al.* 2004). Application of coalescent-theoretic methods (Knowles and Carstens 2007, Rodrigo *et al.* 2008, Zaldívar-Riverón *et al.* 2010, Zhang *et al.* 2011, Nuñez *et al.* 2012, Vuataz *et al.* 2012), the principal of genealogical sorting (Cummings *et al.* 2008), machine learning methods (Bertolazzi *et al.* 2009, Weitschek *et al.* 2013), a heuristic-search-strategy method (O'Meara 2010), Bayesian statistical methods (Yang and Rannala 2010, Zhang *et al.* 2011), and a multimethod 'tip to root' approach (Boykin *et al.* 2012) have been proposed as ways to increase objectivity and biological relevance of species delimitation.

Since decisions regarding species delimitation may affect bioassessment metrics, we attempted to answer three questions: 1) Do different delimitation methods yield different numbers of detectable species? 2) Are differences among methods in the number of species detected associated with the number of haplotypes (i.e., are haplotype-rich species more difficult to delimit, or vice versa)? 3) Do estimates of species richness produced by different delimitation methods differ among sampling sites? To address these questions, we evaluated 4 major classes of species-delimitation criteria (the criteria by which a 'haplotype cluster' is granted a species-level status): genetic distance-based (DB), bootstrap support (BSS), reciprocal monophyly (RM), and coalescent-based (CB). We applied these criteria by constructing two types of phylogenetic trees. First, we constructed neighbor joining trees (NJ), which implement a clustering algorithm that always finds the 'first best' (balanced minimum evolution) tree given the data set. Second, we constructed maximum likelihood trees (ML), which implement an algorithm that heuristically searches a subset of all possible trees to find the highest log-likelihood (lnL) tree. These evaluations focused on three insect genera that are widely encountered in freshwater bioassessment and whose species are difficult to identify

morphologically: Baetis (Ephemeroptera:Baetidae), Eukiefferiella (Diptera:Chironomidae), and Simulium (Diptera:Simuliidae).

Methods

Study Site and Genera

We obtained a subset (2202) of COI sequences from *Baetis*, *Eukiefferiella*, and *Simulium* from a bioassessment study of five streams in the Los Angeles, California (USA), area (Stein *et al.* In press). Benthic macroinvertebrate samples were taken from two reaches at each stream. Expert taxonomists identified specimens morphologically using a standard level of taxonomic effort. These experts identified three distinct species of *Baetis* (*Baetis tricaudatus, Baetis adonis*, and a 3rd unknown, but recognizably distinct, species *Baetis* sp. CA), whereas they identified *Simulium* and *Eukiefferiella* species only to genus. We treated data for each genus separately. The sequences used in our study are publicly available under the BOLD projects CFWIA through CFWIJ (see Table 1 for a complete list of BOLD sample identification codes and Genbank accession numbers).

Sequence Data and Haplotype Collapsing

We selected closely related genera as outgroups for each data set (Table 1). We used a minimum sequence length requirement of 500 basepairs (bp) to reduce uncertainty during NJ and phylogenetic analyses. We translated sequences to amino acids in MEGA (version 5.1; Tamura *et al.* 2011) and aligned them in MUSCLE (version 3.8.31; Edgar 2004). We manually corrected the final alignment so that sequences lacked gaps and consisted of an uninterrupted open-reading-frame, which led us to conclude that no insertions, deletions, or pseudogenes were present in the data sets. Following alignment, we used an open-source, custom Perl script developed for this analysis (dnab_collapse.pl, https://github.com/bpwhite/bioinformatics-toolbox) to reduce the number of individual sequences, and thus, the computational requirements for each analysis. Important information, such as location, haplotype identification, and taxonomic identification were preserved. The end result of this process was sequences that were either unique haplotypes or haplotypes that were present at two or more sites. Each remaining sequence was automatically annotated with the abundance of that haplotype at each location so that

Table 1. Taxa used, number of cytochrome *c* oxidase subunit I (COI) sequences, number of haplotypes for each taxa, outgroups used for each taxa, Barcode of Life Database sample identification numbers (BOLD Sample ID), Genbank accession numbers for each outgroup, nucleotide model used, tree log likelihood (Tree lnL), effective sample size (ESS) of Tree lnL and coalescent BEAST parameters (ESS of Coalescent), null (Yule) model lnL from *gmyc* function, General Mixed Yule Coalescent (GMYC) model lnL, *p* value of the χ^2 Goodness-of-fit test between the null and GMYC model (GMYC *p*), *p* value of the comparison between the single and multiple (Multi *p*), GMYC threshold, and number of putative species detected under the GMYC model (ML Entities).

Taxon	COI Sequences	Haplotypes	Outgroups	BOLD Sample ID	Genbank Accession Number	Nucleotide Model	Tree lnL	ESS of Tree lnL
Baetis	906	104	*Fallceon* sp.	10-SCCWRP-6041	JN297857	HKY+G	–3987.374	203.5
			Centroptilum triangulifer	09LJ17	HM423544			
			Callibaetis ferrugineus	09NBMAY-0053	JQ663249			
Simulium	951	304	*Prosimulium mixtum*	08-SWRC-1082	JF287830	GTR+G	–6018.239	22.5
			Prosimulium travisi	FJ524559	FJ524559			
Eukiefferiella	345	32	*Chironomus riparius*	ChiBlank_F05	HM137935	GTR+I+G	–3826.534	857.1
			Chironomus kiiensis	JQ350720	JQ350720			
			Chironomus jacksoni	CAUS016-09	–			

Taxon	ESS of Coalescent	Null Model lnL	GMYC Model lnL	GMYC *p*	Multi *p*	Threshold (mya)	ML Entities
Baetis	963.9	1195.81	1220.916	<<0.0001	0.884	0.973	3
Simulium	368.4	2258.584	2312.053	<<0.0001	0.917	1.063	7
Eukiefferiella	3744.5	128.73	145.469	<<0.0001	0.999	0.333	10

information about the diversity and abundance of haplotype clusters could be garnered quickly.

Intraspecific Variation Method (NJ+DB)

The DB criterion is based on use of an a priori genetic distance threshold as the cutoff for deciding whether two individuals are members of the same species. This criterion is predicated on the idea that intraspecific genetic variation is small relative to interspecific variation. For example, if the distance cutoff is 2% (Herbert *et al.* 2003, Meyer and Paulay 2005, Rivera and Currie 2009, Sweeney *et al.* 2011), and the calculated genetic distance between individual A and B is 2.5%, then the two individuals are assigned to different species. Variations on this method include use of average intraspecific distances (Hebert *et al.* 2004; Zhou *et al.* 2009, 2011) or variable thresholds depending on the taxa (Sweeney *et al.* 2011). The DB method typically has been applied to NJ trees computed with the algorithm of Saitou and Nei (1987) and the Kimura-2-parameter (K2P; Kimura 1980) measure of genetic distance.

We applied a DB criterion by calculating the nearest-neighbor distance (smallest interspecific distance) between haplotype clusters using the *Species Delimitation* (version 1.04; Masters *et al.* 2011) plugin for Geneious (version 5.6.5; Biomatters, http://www.geneious.com/). Two haplotype clusters that contained a pair of nearest neighbors with >2% K2P distance from each other were considered different putative species.

Statistical Methods (NJ+BSS and ML+RM)

BSS is the proportion of bootstrap replicates in which particular sequences clustered together when the NJ algorithm is applied (Felsenstein 1985). For example, if a node achieves 95% bootstrap support, then that node and all of its children were grouped together in 95% of the bootstrap replicates. This method has been used in large-scale DNA barcoding studies by Zhou *et al.* (2009, 2011), Mecklenburg *et al.* (2011), and Lakra *et al.* (2011).

We implemented the NJ+BSS method using 1000 bootstrap replicates in MEGA. We used K2P distance because it is considered 'standard' in DNA barcode studies (Herbert *et al.* 2003, 2004; Zhou *et al.* 2009, 2011; Ocegura-Figueroa *et al.* 2010; Sweeney *et al.* 2011). However, Srivathsan and Meier (2011) and Collins *et al.* (2012) recently suggested that K2P is rarely the best nucleotide model for COI-only data sets. We applied the BSS criterion to the bootstrapped NJ tree to define putative species based on a bootstrap support cutoff of 95%.

RM is a statistical approach based on the principal that individuals from different species will separate consistently into distinct monophyletic clades with >95% statistical support. RM can be applied to either maximum parsimony (MP) or ML trees, but a more rigorous test of monophyly (not done here) requires that the observed branching pattern be tested against a random branching pattern (Rosenberg 2007).

We implemented the ML+RM method by first identifying the optimal nucleotide model for each data set with jModelTest (version 0.1.1; Posada 2008). Following nucleotide model selection, we constructed ML phylogenetic trees using a Bayesian phylogenetic program, BEAST (version 1.7.4; Drummond *et al.* 2012) with a coalescent-tree prior and 3 different molecular clock models: strict, relaxed lognormal, and relaxed exponential. Each clock model began with a normally distributed clock rate with a mean of 0.02 substitutions/million y (s/my) (Brown *et al.* 1979) and a standard deviation of 0.005 s/my. We ran Monte Carlo Markov Chain (MCMC) simulation for 10 million steps and sampled trees from the MCMC at 1000-step intervals. We checked parameter values for effective sample sizes (ESS) >200 and convergence by plotting marginal probabilities in Tracer (version 1.5; Drummond *et al.* 2012). We discarded the first 20% of trees sampled as burn-ins. We loaded the remaining 8001 trees into TreeAnnotator (version 1.7.4; Drummond *et al.* 2012) for construction of the maximum clade credibility (MCC) tree and calculation of posterior probabilities and node ages. After trees were annotated, we used a Bayes factor (BF) analysis to test whether the data were clock-like (Drummond and Rambaut 2007). In this analysis, the marginal likelihoods of each tree are estimated using the harmonic mean, and the possible improvement of one model over another is assessed by dividing their marginal likelihoods when those models differ by only one parameter (in this case, the clock model). The ratio of this division is the BF. An improvement of one model over another is considered significant if BF >2. Following selection of the best tree clock model, the MCC was annotated in FigTree (version 1.4; http://tree.bio.ed.ac.uk/software/figtree/).

Modeling Method (ML+CB)

CB is a modeling approach derived from population genetics and is based on the principal that individuals that possess different species-level coalescent points (hypothetical ancestors of haplotypes, alleles, or species from which point all current members of a population were descended) are members of different species. Here we consider the COI gene tree to be analogous to the species tree, but in many cases, gene trees do not match species trees (Liu and Pearl 2007). Analysis of multiple genes typically is required to obtain an accurate species coalescent point.

A custom R script was created (dnab_coalesce.r, also available from: https://github.com/bpwhite/bioinformatics-toolbox) to run the CB species-delimitation analysis. This script makes use of the *splits* package (http://r-forge.r-project.org/projects/splits/), and imports the resultant MCC trees for each data set into the General Mixed Yule Coalescent (GMYC) function (*gmyc*). The *gmyc* function finds the ML threshold for the transition from a Yule process (interspecific branching rates) to a coalescent process (intraspecific branching rates; Pons *et al.* 2006, Fontaneto *et al.* 2007, Knowles and Carstens 2007, Monaghan *et al.* 2009, Nuñez *et al.* 2012, Vuataz *et al.* 2012). A likelihood ratio test is automatically performed to compare the coalescent model to a Yule model of evolution, and if the ratio results in a p-value <0.05, the coalescent model is accepted over the Yule model and the putative species entities can be considered statistically significant. We ran the GMYC model test for a single ML threshold and multiple ML thresholds, wherein the threshold was allowed to vary across lineages. We compared the results of the two models with a χ^2 Goodness-of-Fit test (also available in the *splits* package under the function *compare*. The multiple threshold test was considered an improvement over the single threshold test if the data fit that model significantly better (p <0.05). After model selection, the dnab_coalesce.r script outputs the resultant species delimitation using the *spec.list* function into a comma-separated-value (CSV) format for import into other programs.

Species Richness and Abundance Calculations

We assigned four putative species identifications (one for each method) to each individual sequence and used a χ^2 Goodness-of-Fit test to assess whether species richness was affected by the method used. We assessed the effects of species-delimitation methods on stream species richness by summing the number of putative species encountered in each stream for each method. We assessed where shifts in the abundance of species might occur by summing the number of individuals given a particular species identification for each method.

Results

Data reduction of the 3 data sets decreased the numbers of sequences from 951 to 201 for *Baetis*, 906 to 389 for *Simulium*, and 345 to 32 for *Eukiefferiella*. The number of putative species did not differ among delimitation methods for any genus ($\chi^2_{\text{Table 2}} = 1.327$, $p > 0.05$).

Intraspecific Variation Method (NJ+DB)

The NJ+DB method delimited more putative *Baetis* species than all other methods (Table 2) because it split *Baetis* 1 into 2 species (1 and 2; Figure 1). The NJ+DB method delimited fewer putative *Eukiefferiella* species than all other methods (Table 2) because it lumped *Eukiefferiella* 5 and 6 into 1 species. The genetic distance between *Eukiefferiella* 5 and 6 was 1.9%, thus species 6 missed the cutoff by 0.1% (Figure 2). The NJ+DB method delimited more putative *Simulium* species than all other methods (Table 2) because it split *Simulium* 1 into 2 species (1 and 2) and *Simulium* 9 into 2 species (9 and 10; Figure 3).

Statistical Methods (NJ+BSS and ML+RM)

Both statistical based methods resulted in identical species delimitations in all three genera, but the use of ML tree construction methods increased support values for many nodes over the bootstrap support values (Figures 1 through 3).

Table 2. Number of species identified in each genus with each species-delimitation method. ML = maximum likelihood; CB = coalescent-based; RM = reciprocal monophyly; NJ = neighbor-joining; BSS = bootstrap support; and DB = distance-based.

Taxon	ML+CB	ML+RM	NJ+BSS	NJ+DB
Baetis	3	4	4	5
Eukiefferiella	10	10	10	9
Simulium	8	7	7	10

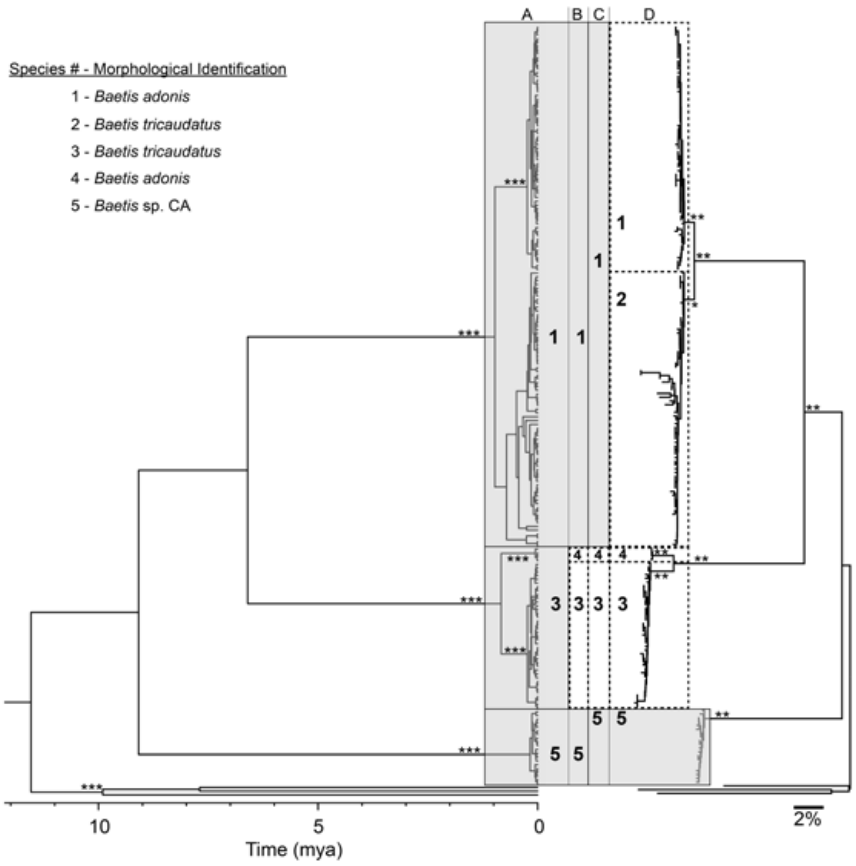

Figure 1. Maximum likelihood, strict-clock tree of 201 *Baetis* cytochrome *c* oxidase subunit I (COI) sequences computed using BEAST with a coalescent tree prior and the HKY+G nucleotide model; species entities were delimited using the *gmyc* function of the SPLITS package 1.0-14 (left). Neighbor-joining tree computed using MEGA 5.05 and the Kimura 2-parameter distance nucleotide model (right). Putative species are numbered in order of appearance on the tree and highlighted in shaded grey boxes. Putative species that undergo splitting are indicated by dashed boxes. A = ML+CB; B = ML+RM; C = NJ+BSS; D = NJ+DB; * = 0.80 - 0.94 node support; ** = 0.95 - 0.99 node support; and * = 1.00 node support.**

Modeling Method (ML+CB)

Each data set had a different nucleotide model. Hasegawa-Kishino-Yano + gamma [HKY+G] was selected for *Baetis*, general time reversible + gamma (GTR+G) was selected for *Simulium*, and general time reversible + invariant + gamma (GTR+I+G) was selected for *Eukiefferiella* (Table 1). The lognormal and exponential relaxed clock models were not significant improvements over the strict clock model for any genus (BF <2 in all cases), so the strict clock model was used for both *Baetis* and *Eukiefferiella*. In the case of *Simulium*, negative branch lengths in the strict clock tree made the application of the GMYC model impossible. We used the lognormal clock tree instead because it had only a slightly faster mean rate than the strict clock tree (strict: 0.199 vs lognormal: 0.244). The *Simulium* MCMC may have been undersampled because the ESS for the likelihood parameter was <200, whereas the coalescent parameter was >200 (Table 1). For all three genera, the GMYC model was selected over the null model of evolution (Table 1), and the multiple ML threshold model was not a significant improvement over the single threshold model (Table 1). The ML+CB method produced the same number of species as the ML+RM and NJ+BSS methods, but the designations of those species were different, for example, *Baetis* 3 was not split into *Baetis* 3 and 4 (Figure 1), whereas it was in the other 3 methods. Moreover, *Simulium* 8 and 9 were identified by the ML+CB method, but *Simulium* 10 was not (Figure 3).

Shifts in Species Abundance and Richness

The only difference in species abundances among delimitation methods was for *Baetis* 1 and 2. The NJ+DB method yielded 2 species consisting of 364

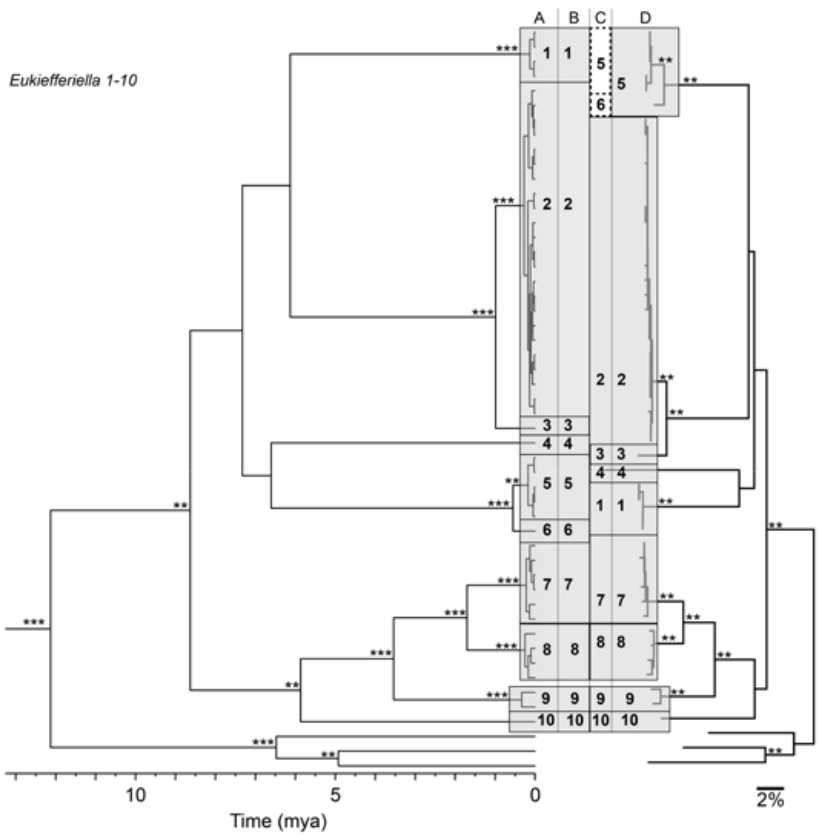

Figure 2. Maximum likelihood, strict-clock tree of 32 *Eukiefferiella* cytochrome *c* oxidase subunit I (COI) sequences computed using BEAST with a coalescent tree prior and the GTR+I+G nucleotide model; species entities were delimited using the *gmyc* function of the SPLITS package 1.0-14 (left). Neighbor-joining tree computed using MEGA 5.05 and the Kimura 2-parameter distance nucleotide model (right). Putative species are numbered in order of appearance on the tree and highlighted in shaded grey boxes. Putative species that undergo splitting are indicated by dashed boxes. A= ML+CB; B = ML+RM; C = NJ+BSS; D = NJ+DB; * = 0.80 - 0.94 node support; ** = 0.95 - 0.99 node support; and * = 1.00 node support.**

and 429 individuals, whereas the other methods yielded 1 species with 793 individuals. Minor differences in species abundances among methods were present in all three genera, but these differences were limited mostly to the presence or absence of a few rare haplotypes. Species richness differed consistently among methods and sites (Figure 4). When estimates differed among methods, the NJ+DB method typically produced higher species richness than the other methods. In all cases the RM and BSS method produced identical abundance and richness estimates.

Discussion

Four species-delimitation methods applied to a data set of 2202 COI sequences from 3 genera of insect larvae from southern California yielded similar estimates of species richness. Where standard morphological identification effort yielded 5 distinct taxa (*Baetis adonis*, *Baetis tricaudatus*, *Baetis sp. CA, Simulium*, and *Eukiefferiella*), DNA barcodes yielded 19 to 25 putative species, a 4x increase in resolution over the standard level of identification. The differences among species delimitations, although not statistically significant, tended to be associated with abundant and diverse taxa. This result suggests that the uncertainty associated with species delimitations derived from DNA barcoding does not arise from the algorithm used, but is a byproduct of the inherent limitations of COI as a species-level phylogenetic marker. Differences among delimitation methods are not likely to result in large changes in bioassessment metric scores based on taxon richness.

The absence of noticeable differences in species abundances under different delimitation methods (except *Baetis* 1 and 2) suggests that richness metrics that take abundances into account might be unaffected by different delimitation methods,

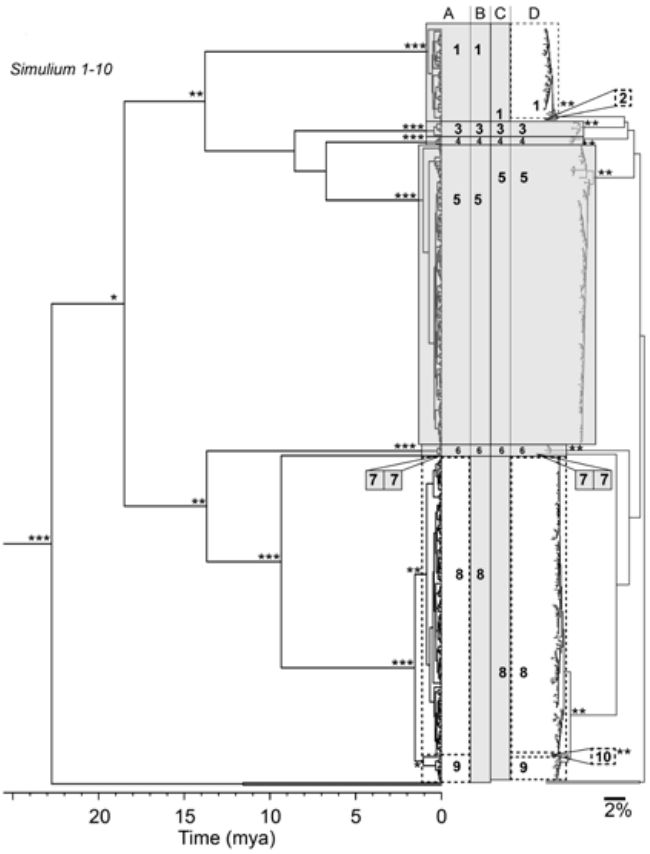

Figure 3. Maximum likelihood, lognormal-clock tree of 389 *Simulium* cytochrome *c* oxidase subunit I (COI) sequences computed using BEAST with a coalescent tree prior and the GTR+G nucleotide model; species entities were delimited using the *gmyc* function of the SPLITS package 1.0-14 (left). Neighbor-joining tree computed using MEGA 5.05 and the Kimura 2-parameter distance nucleotide model (right). Putative species are numbered in order of appearance on the tree and highlighted in shaded grey boxes. Putative species that undergo splitting are indicated by dashed boxes. A= ML+CB; B = ML+RM; C = NJ+BSS; D = NJ+DB; * = 0.80 - 0.94 node support; ** = 0.95 - 0.99 node support; and * = 1.00 node support.**

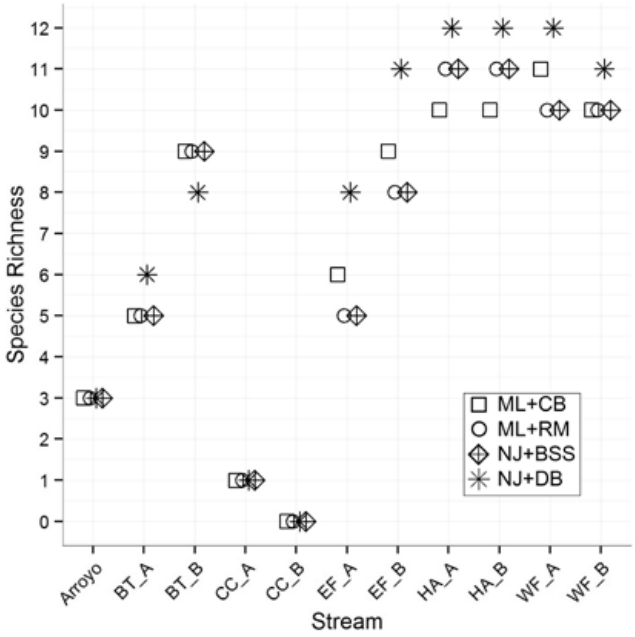

Figure 4. Differences in species richness observed at different southern California stream reaches.

whereas presence-absence-type richness metrics may be more directly influenced by even subtle shifts in species designations (e.g., *Simulium* 2 and 10 exist only under the NJ+DB method and have extremely low abundances [1 and 2 individuals, respectively]). However, even those small differences may not be large enough to significantly affect bioassessment metrics beyond the greater taxonomic resolution already provided by DNA barcoding.

Species delimitations that differed among methods tended to be associated with high-abundance species with many different COI haplotypes, a potentially important trend warranting further examination. For example, putative species that experienced splits under the NJ+DB method (*Baetis* 1-2, 3-4, and *Simulium* 1-2, 9-10) were usually very abundant (>30 individuals encountered). Large populations tend to be less prone to chance events like genetic drift that eliminate rare haplotypes (given enough time), it seems plausible that very abundant species will maintain more rare haplotypes in its gene pool than will less abundant species. This idea received support from Bergsten *et al.* (2012), who concluded that the uncertainty of species identifications for *Agabini* diving beetles increased significantly when sampled over increasing geographic distances. They found that a sample size of 70 individuals was necessary to capture 95% of intraspecific diversity. We observed a pattern in *Baetis* sp. 1 and 2, which made up 83% of *Baetis* encountered (793 individuals), of increased diversity with increased intraspecific sampling effort, a result that further supports Bergsten's *et al.* (2012) hypothesis. In our data set, *Baetis* 1 and 2 had an average intraspecific K2P distance of 1.6% and a maximum pairwise K2P distance of 5.1% when lumped together according to the NJ+BS, ML+RM, and ML+CB criteria. The less frequently encountered *Baetis* species (3, 4, and 5), which made up only 17% of *Baetis* encountered (158 individuals), did not exhibit intraspecific pairwise distances >1%. These differences in sampling effort reflect natural abundances of these species and not targeted sampling effort toward one species or another. In

contrast, we also found many individuals in *Baetis* 1 and 2 that were separated by great geographic distances but shared identical haplotypes. For example, 8 individuals found in Conejo Creek had haplotypes identical to those 50 individuals in Big Tujunga Wash. The 2 streams are separated by a geographic distance of 64 km and an elevation difference of 335 m. We did not see more than one or two instances of shared haplotypes over great distances in *Simulium* or *Eukiefferiella*.

The multimodal pattern of genetic variation within the *Baetis* 1 and 2 complex consists of a mixture of low and high diversity over broad distances. Such a pattern might be explained if these species were parthenogenetic. Many mayfly species exhibit parthenogenesis (Bergman and Hilsenhoff 1978; McCafferty and Morihara 1979; Funk *et al.* 2006, 2008), and Funk *et al.* (2010) suggested that most, if not all, mayflies may be facultatively parthenogenetic. Females may reproduce parthenogenetically, but offspring can be either male or female and can readily revert to sexual reproduction. Vanoverbeke and De Meester (1997) found no relationship between geographic distance and genetic distance in parthenogenetically reproducing populations of cladoceran branchiopods (*Daphnia magna*), a result similar to our finding of shared haplotypes over large distances (64 km). However, we cannot rule out the possibility that this pattern is a by-product of normal metazoan mitochondrial inheritance and that shared haplotypes are a consequence of evolutionarily recent dispersal events.

The increased taxonomic resolution offered by DNA barcoding could be used to understand better the life histories of the seemingly cryptic *Baetis* observed in our study, which might in turn help researchers use traits-based approaches to ecology and improve future bioassessment tools (Verberk *et al.* 2013). In the case of *Baetis* spp. in southern California, the addition of nuclear loci to DNA barcode data sets might help researchers distinguish between normal and parthenogenetic modes of inheritance (Buckley *et al.* 2008). When systematists undertake taxonomic revisions of morphologically cryptic species in light of molecular data (e.g., in the *Atyaephyra* genus of freshwater shrimp; Christodoulou *et al.* 2012), specific life-history traits, such as parthenogenesis, could be included in descriptions, transferred to traits databases, and associated with DNA barcodes.

Coalescent models take into account the natural birth and death processes of populations, avoid the use of *a priori* distance cutoffs (e.g., the 2% cutoff), and provide a statistical framework for testing species delimitations. Furthermore, the GMYC model provides estimates of the times of transitions from inter- to intraspecies branching patterns and effectively links the fields of population genetics and phylogenetics. However, multiple genes from both nuclear and mitochondrial DNA are required to obtain a robust estimate of the species coalescent (Heled and Drummond 2010, Fujita *et al.* 2012). If we regard a species as any separately evolving metapopulation of lineages (de Quieroz 2007), the ML+CB method stands beside distance and monophyly-based methods as a separate and unique line of evidence in the diagnosis of distinct lineages. Thus, combining the results of several species-delimitation methods might allow researchers to draw confident conclusions when diagnosing a lineage (as in Boykin *et al.* 2012).

Differentiating between cases of rare, divergent haplotypes within a species and cryptic species is one of the greatest challenges for users of DNA barcodes as species-level markers. This challenge probably will be overcome only by using multigene and multilocus reference libraries. A combination of mitochondrial genes (e.g., COI, CYTB, and 16S) and nuclear genes (e.g., 18S, 28S, ITS1, and ITS2) would allow robust estimates of species coalescence and improved phylogenetic resolution. Use of multigene data sets in routine bioassessments might be cost prohibitive because linking multiple genes to a single voucher specimen requires sorting and molecular tagging of each individual voucher specimen. As metabarcoding of environmental samples becomes more prevalent (Ji *et al.* 2013, Carew *et al.* 2013), sorting of individual organisms might become less problematic. One solution might be to maintain multigene reference libraries of local taxa identified with rigorous species-delimitation methods, but to use single-gene methods during routine field sampling (e.g., COI- or 16S-based DNA barcoding coupled with distance-based algorithms). Multigene reference libraries would allow researchers to know a priori when instances of mitochondrial introgression (incorrect lumping) or ancestral polymorphisms (incorrect splitting) might produce spurious single-gene delimitations so that degrees of confidence could be assigned to molecularly derived species delimitations.

Literature Cited

Ball, S.L., P.D.N. Hebert, S.K. Burian and J.M. Webb. 2005. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *Journal of the North American Benthological Society* 24:508-524.

Bergman, E.A. and W.L. Hilsenhoff. 1978. Parthenogenesis in the mayfly genus *Baetis* (Ephemeroptera: Baetidae). *Annals of the Entomological Society of America* 71:167-168.

Bergsten, J., D.T. Bilton, T. Fujisawa, M. Elliott, M.T. Monaghan, M. Balke and A.P. Vogler. 2012. The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology* 61:851-869.

Bertolazzi, P., G. Felici and E. Weitschek. 2009. Learning to classify species with barcodes. BMC *Bioinformatics* 10:S7.

Bickford, D., D.J. Lohman, N.S. Sodhi, P.K. Ng, R. Meier, K. Winker, K.K. Ingram and I. Das. 2007. Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution* 22:148-155.

Boykin, L.M., K.F. Armstrong, L. Kubatko and P.D. Barro. 2012. Species delimitation and global biosecurity. *Evolutionary Bioinformatics* 8:1-37.

Brown, W.M., M. George and A.C. Wilson. 1979. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences* 76:1967-1971.

Buckley, T.R., D. Attanayake, D. Park, S. Ravindran, T.R. Jewell and B.B. Normark. 2008. Investigating hybridization in the parthenogenetic New Zealand stick insect *Acanthoxyla* (Phasmatodea) using single-copy nuclear loci. *Molecular Phylogenetics and Evolution* 48:335-349.

Carew, M.E., V.J. Pettigrove, L. Metzeling and A.A. Hoffmann. 2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology* 10:1-15.

Christodoulou, M., A. Antoniou, A. Magoulas and A Koukouras. 2012. Revision of the freshwater genus *Atyaephyra* (Crustacea, Decapoda, Atyidae) based on morphological and molecular data. *ZooKeys* 229:53-110.

Collins, R.A., L.M. Boykin, R.H. Cruickshank and K.F. Armstrong. 2012. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution* 3:457-465.

Cummings, M.P., M.C. Neel and K.L. Shaw. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* 62:2411-2422.

Dawnay, N., R. Ogden, R. McEwing, G.R. Carvalho and R.S. Thorpe. 2007. Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Science International* 173:1-6.

de Queiroz, K. 2007. Species concepts and species delimitation. *Systematic Biology* 56:879-886.

Drummond, A.J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7:214.

Drummond, A.J., M.A. Suchard, D. Xie and A. Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biolology and Evolution* 29:1969-1973.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792-1797.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Fontaneto, D., E.A. Herniou, C. Boschetti, M. Caprioli, G. Melone, C. Ricci and T.G. Barraclough. 2007. Independently evolving species in asexual bdelloid rotifers. *PLOS Biology* 5:e87.

Fujita, M.K., A.D. Leache, F.T. Burbrink, J.A. McGuire and C. Moritz. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology and Evolution* 27:480-488.

Funk, D.H., J.K. Jackson and B.W. Sweeney. 2006. Taxonomy and genetics of the parthenogenetic mayfly *Centroptilum triangulifer* and its sexual sister *Centroptilum alamance* (Ephemeroptera:Baetidae). *Journal of the North American Benthological Society* 25:417-429.

Funk, D.H., J.K. Jackson and B.W. Sweeney. 2008. A new parthenogenetic mayfly (Ephemeroptera:E phemerellidae:*Eurylophella* Tiensuu) oviposits by

abdominal bursting in the subimago. *Journal of the North American Benthological Society* 27:269-279.

Funk, D.H., B.W. Sweeney and J.K. Jackson. 2010. Why stream mayflies can reproduce without males but remain bisexual: a case of lost genetic variation. *Journal of the North American Benthological Society* 29:1258-1266.

Hajibabaei, M., D.H. Janzen, J.M. Burns, W. Hallwachs and P.D.N. Hebert. 2006. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America* 103:968-971.

Hebert, P.D.N., A. Cywinska, S.L. Ball and J.R. deWaard. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences* 270:313-321.

Hebert, P.D.N., M.Y. Stoeckle, T.S. Zemlak and C.M. Francis. 2004. Identification of birds through DNA barcodes. *PLoS Biology* 2:e312.

Heled, J. and A.J. Drummond. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* 27:570-580.

Janzen, D.H., W. Hallwachs, P. Blandin, J.M. Burns, J. Cadiou, I. Chacon, T. Dapkey, A.R. Deans, M.E. Epstein, B. Espinoza, J.G. Franclemont, W.A. Haber, M. Hajibabaei, J.P.W. Hall, P.D.N. Hebert, I.D. Gauld, D.J. Harvey, A. Hausmann, I.J. Kitching, D. Lafontaine, J.-F. Landry, C. Lemaire, J.Y. Miller, J.S. Miller, L. Miller, S.E. Miller, J. Montero, E. Munroe, S.R. Green, S. Ratnasingham, J.E. Rawlins, R.K. Robbins, J.J. Rodriguez, R. Rougerie, M.J. Sharkey, M.A. Smith, M.A. Solis, J.B. Sullivan, P. Thiaucourt, D.B. Wahl, S.J. Weller, J.B. Whitfield, K.R. Willmott, D.M. Wood, N.E. Woodley and J.J. Wilson. 2009. Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. *Molecular Ecology Resources* 9:1-26.

Ji, Y., L. Ashton, S.M. Pedley, D. P. Edwards, Y. Tang, A. Nakamura, R. Kitching, P.M. Dolman, P. Woodcock, F.A. Edwards, T.H. Larsen, W.W. Hsu, S. Benedick, K.C. Hamer, D.S. Wilcove, C. Bruce, X. Wang, T. Levi, M. Lott, B.C. Emerson and D.W. Yu. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters* 16:1245-1257.

Jones, F.C. 2008. Taxonomic sufficiency: The influence of taxonomic resolution on freshwater bioassessments using benthic macroinvertebrates. *Environmental Reviews* 16:45-69.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111-120.

Knowles, L.L. and B.C. Carstens. 2007. Delimiting species without monophyletic gene trees. *Systematic Biology* 56:887-895.

Lakra, W.S., M.S. Verma, M. Goswami, K.K. Lal, V. Mohindra, P. Punia, A. Gopalakrishnan, K.V. Singh, R.D. Ward and P. Hebert. 2011. DNA barcoding Indian marine fishes. *Molecular Ecology Resources* 11:60-71.

Liu, L. and D.K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56:504-514.

Maralit, B.A., R.D. Aguila, M.F.H. Ventolero, S.K.L. Perez and M.D. Santos. 2013. Detection of mislabeled commercial fishery by-products in the Philippines using DNA barcodes and its implications to food traceability and safety. *Food Control* 33:119-125.

Masters, B.C., V. Fan and H.A. Ross. 2011. Species delimitation-a Geneious plugin for the exploration of species boundaries. *Molecular Ecolology Resources* 11:154-157.

McCafferty, W.P. and D.K. Morihara. 1979. The male of *Baetis macdunnoughi* Ide and notes on parthenogenetic populations within *Baetis* (Ephemeroptera: Baetidae). *Entomological News* 90:26-28.

Mecklenburg, C.W., P.R. Møller and D. Steinke. 2011. Biodiversity of arctic marine fishes: taxonomy and zoogeography. *Marine Biodiversity* 41:109-140.

Meyer, C.P. and G. Paulay. 2005. DNA barcoding: error rates based on comprehensive sampling. *PLOS Biology* 3:e422.

Monaghan, M.T., R. Wild, M. Elliot, T. Fujisawa, M. Balke, D.J. Inward, D.C. Lees, R. Ranaivosolo, P. Eggleton, T.G. Barraclough and A.P. Vogler. 2009. Accelerated species inventory on Madagascar using

coalescent-based models of species delineation. *Systematic Biology* 58:298-311.

Nuñez, J.J., A. Vejar-Pardo, B.E. Guzmán, E.H. Barriga and C.S. Gallardo. 2012. Phylogenetic and mixed Yule-coalescent analyses reveal cryptic lineages within two South American marine snails of the genus *Crepipatella* (Gastropoda:Calyptraeidae). *Invertebrate Biology* 131:301-311.

Oceguera-Figueroa, A., V. León-Règagnon and M.E. Siddall. 2010. DNA barcoding reveals Mexican diversity within the freshwater leech genus *Helobdella* (Annelida: Glossiphoniidae). *Mitochondrial DNA* 21:24-29.

O'Meara, B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology* 59:59-73.

Pilgrim, E.M., S.A. Jackson, S. Swenson, I. Turcasanyi, E. Friedman, L. Weigt and M.J. Bagley. 2011. Incorporation of DNA barcoding into large-scale biomonitoring program: opportunities and pitfalls. *Journal of the North American Benthological Society* 30:217-231.

Pons, J., T.G. Barraclough, J. Gomez-Zurita, A. Cardoso, D.P. Duran, S. Hazell and A.P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology* 55:595-609.

Posada, D. 2008. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* 25:1253-1256.

Ratnasingham, S. and P.D. Hebert. 2007. BOLD: the Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes* 7:355-364.

Rivera, J. and D.C. Currie. 2009. Identification of Nearctic black flies using DNA barcodes (Diptera: Simuliidae). *Molecular Ecology Resources* 9:224-236.

Rodrigo, A., F. Bertels, J. Heled, R. Noder, H. Shearman and P. Tsai. 2008. The perils of plenty: What are we going to do with all these genes? *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 363:3893-3902.

Rosenberg, N.A. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution* 61:317-323.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406-425.

Shendure, J. and H. Ji. 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26:1135-1145.

Srivathsan, A. and R. Meier. 2011. On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics* 28:190-194.

Stein, E.D., B.P. White, R.D. Mazor, J.K. Jackson, J.M. Battle, P.E. Miller, E.M. Pilgrim and B.W. Sweeney. In press. Does DNA barcoding improve performance of traditional stream bioassessment metrics? *Freshwater Science*.

Stein, E.D., B.P. White, R.D. Mazor, P.E. Miller and E.M. Pilgrim. 2013. Evaluating ethanol-based sample preservation to facilitate use of DNA barcoding in routine freshwater biomonitoring programs using benthic macroinvertebrates. *PLOS ONE* 8:e51273.

Sweeney, B.W., J.M. Battle, J.K. Jackson and T. Dapkey. 2011. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *Journal of the North American Benthological Society* 30:195-216.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28:2731-2739.

Valentini, A., F. Pompanon and P. Taberlet. 2009. DNA barcoding for ecologists. *Trends in Ecology and Evolution* 24:110-117.

Vanoverbeke, J. and L. De Meester. 1997. Among-populational genetic differentiation in the cyclical parthenogen *Daphnia magna* (Crustacea, Anomopoda) and its relation to geographic distance and clonal diversity. *Hydrobiologia* 360:135-142.

Verberk, W.C.E.P., C.G.E. van Noordwijk and A.G. Hildrew. 2013. Delivering on a promise: Integrating

species traits to transform descriptive community ecology into a predictive science. *Freshwater Science* 32:531-547.

Vuataz, L., M. Sartori, J.L. Gattolliat and M.T. Monaghan. 2012. Endemism and diversification in freshwater insects of Madagascar revealed by coalescent and phylogenetic analysis of museum and field collections. *Molecular Phylogenetics and Evolution* 66:979-991.

Weitschek, E., R. Velzen, G. Felici and P. Bertolazzi. 2013. BLOG 2.0: A software system for character-based species classification with DNA Barcode sequences. What it does, how to use it. *Molecular Ecology Resources* 13:1043-1046.

Yang, Z. and B. Rannala. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 107:9264-9269.

Zaldívar-Riverón, A., J.J. Martínez, F.S. Ceccarelli, V.S. De Jesús-Bonilla, A.C. Rodríguez-Pérez, A. Reséndiz-Flores and M.A. Smith. 2010. DNA barcoding a highly diverse group of parasitoid wasps (Braconidae: Doryctinae) from a Mexican nature reserve. *Mitochondrial DNA* 21:18-23.

Zhang, C., D. X. Zhang, T. Zhu and Z. Yang. 2011. Evaluation of a Bayesian coalescent method of species delimitation. *Systematic Biology* 60:747-761.

Zhou X., S.J. Adamowicz, L.M. Jacobus, E. DeWalt and P.D.N. Hebert. 2009. Toward a comprehensive barcode library for Arctic life-Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. *Frontiers in Zoology* 6:30.

Zhou, X., J.L. Robinson, C.J. Geraci, C.R. Parker, O.S. Flint, D. Etnier, D. Ruiter, R.E. DeWalt, L.M. Jacobus and P.D.N. Hebert. 2011. Accelerated construction of a regional DNA barcode reference library: Caddisflies (Trichoptera) in the Great Smoky Mountains National Park. *Journal of the North American Benthological Society* 30:131-162.

Acknowledgements

We thank Tomochika Fujisawa for providing an advance release copy of the Species List script essential to the coalescent delimitation analysis. We are also thankful to Daniel Pickard for providing taxonomic expertise and valuable feedback on *Baetis* species identifications. We are grateful to the Canadian Centre for DNA Barcoding, which provided sequencing with support from the Life Technologies Corporation. We also thank the two anonymous referees for their feedback which greatly improved the quality of the manuscript.